DOCUMENT RESUME

ED 440 149 TM 030 761

AUTHOR Stecher, Brian M.; Barron, Sheila I.
TITLE Quadrennial Milepost Accountability Testing in Kentucky. CSE Technical Report.
INSTITUTION California Univ., Los Angeles. Center for the Study of Evaluation.; Center for Research on Evaluation, Standards, and Student Testing, Los Angeles, CA.
SPONS AGENCY Office of Educational Research and Improvement (ED), Washington, DC.
REPORT NO CSE-TR-505
PUB DATE 1999-06-00
NOTE 40p.
CONTRACT R305B60002
PUB TYPE Reports - Research (143)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Academic Standards; *Accountability; Behavior Patterns; Case Studies; Elementary Secondary Education; Models; State Programs; *State Standards; Surveys; *Teachers; *Teaching Methods; Test Use; *Testing Programs
IDENTIFIERS *Kentucky; Kentucky Instructional Results Information System

ABSTRACT

Kentucky has been implementing test-based accountability for almost a decade, making it a good site for studying the effects of the milepost testing model. In 1996, a study was undertaken of the impact of standards-based assessment on classroom practices in Kentucky. Kentucky teachers (n=365) were surveyed about their classroom practices and other school practices during the 1996-1997 and 1997-1998 school years. Case studies of a small group of exemplary teachers were also conducted. The 1997-1998 survey involved writing and mathematics teachers, both subjects that were assessed using portfolios. The study confirms some of the positive effects of test-based accountability that have been reported previously, but it also reveals previously unexamined negative consequences arising from high-stakes tests. On the positive side, teachers reacted to the Kentucky Instructional Results Information System by changing their behaviors in ways that were consistent with the specific targets of the system. On the negative side, teachers focused on the most proximal aspects of the system (tests) rather than the more distant goals. The testing and accountability system may be leading teachers to a near-sightedness with several consequences. One is large swings in exposure to specific subjects from year to year. The focus on testing and on milepost grade levels may deflect attention from the cumulative nature of education. (Contains 15 tables and 22 references.) (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

CRESST

National Center for Research on Evaluation, Standards, and Student Testing

ED 440 149

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Quadrennial Milepost Accountability Testing in Kentucky

CSE Technical Report 505

Brian M. Stecher and Sheila I. Barron
CRESST/RAND Education

TM030761

UCLA Center for the Study of Evaluation

In Collaboration With

BEST COPY AVAILABLE

Quadrennial Milepost Accountability Testing in Kentucky

CSE Technical Report 505

Brian M. Stecher and Sheila I. Barron
CRESST/RAND Education

June 1999

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
Center for the Study of Evaluation (CSE)
Graduate School of Education & Information Studies
University of California, Los Angeles
Los Angeles, CA 90095-1522
(310) 206-1532

Project 1.5 The Effects of Standards-Based Assessments on Schools and Classrooms, National Center for Research on Evaluation, Standards and Student Testing (CRESST), RAND Education Brian M. Stecher, Senior Researcher

Copyright © 1999 The Regents of the University of California

The work reported herein was supported under the Educational Research and Development Centers Program, PR/Award Number R305B60002, as administered by the Office of Educational Research and Improvement, U. S. Department of Education.

The findings and opinions expressed in this report do not reflect the positions or policies of the National Institute on Student Achievement, Curriculum, and Assessment, the Office of Educational Research and Improvement, or the U. S. Department of Education.

QUADRENNIAL MILEPOST ACCOUNTABILITY TESTING IN KENTUCKY[1]

Brian M. Stecher and Sheila I. Barron
National Center for Research on Evaluation, Standards and Student Testing (CRESST)
RAND Education

Abstract

Kentucky provides an opportunity to study a high-stakes test-based accountability system that uses milepost testing to see how schools react to the accountability pressures. Kentucky has been in the forefront of the test-based accountability movement, and many states have looked to Kentucky when designing their own accountability systems. Hence, lessons learned in Kentucky will be immediately relevant to other states.

Background

The past two decades have seen a dramatic increase in states' use of tests as educational policy tools, although researchers have raised questions about the validity of scores produced by high stakes state tests and the impact of these tests on classroom practices. These questions remain unresolved, in part, because the increase in state testing has occurred rapidly and testing practices have changed as well. The number of states with mandated student testing programs grew from 29 in 1980 to 46 in 1992 (Office of Technology Assessment, 1992). The

[1] This project would not have been possible without assistance from the Kentucky Department of Education and cooperation from teachers across the state. In particular, we want to acknowledge the support of the staff from the Kentucky Department of Education, including Brian Gong, Sue Rigney, Starr Lewis, and Jonathan Dings. In addition, we would like to thank the hundreds of Kentucky classroom teachers who took the time to complete our survey.

Our RAND colleagues Susan Weinblatt, Suzanne Perry, and Linda Daly deserve credit for coordinating the statewide survey effort, including production, distribution, monitoring, review, and data editing. Our thanks to Cathy Krop for assisting with survey design and Tammi Chun for coding the open-response questions.

increase in state testing has been accompanied by the introduction of new types of assessments, an increase in the stakes attached to scores, and the incorporation of tests into formal accountability systems. A recent survey of state assessment practices found that 39 states were administering some form of performance assessment and six others were planning or developing performance assessments; 24 states attached stakes to their tests in the form of student recognition, promotion, or graduation; and 40 states used test scores for school accountability purposes (Bond et al., 1995).

The growth in state testing has been motivated by two broad goals: to produce valid indicators of student and school outcomes and to promote improved instruction. However, research suggests that state tests, particularly those with high stakes, do not always achieve these goals. High-stakes state testing programs, even those employing multiple choice tests, do not always produce valid information on students or schools (Linn, Graue, & Saunders, 1990; Koretz, Linn, Dunbar, & Shepard, 1991). Similarly, high stakes testing can have undesirable effects on instructional practice, most notably narrowing of the curriculum and undue focus on test-like activities (Kellaghan & Madaus, 1991; Shepard & Dougherty, 1991; Smith & Rothenberg, 1991).

The use of performance tasks, particularly portfolios, exacerbates the problem of score validity (Koretz, 1998). For example, studies of portfolio-based assessments in Vermont and Kentucky found that student work could not be rated reliably and that scores were not valid for their intended purposes (Koretz, Klein, McCaffrey, & Stecher, 1993; Hambleton et al., 1995). More recent studies have also raised questions about the validity of scores from on-demand open-response testing, as well (Koretz & Barron, 1998).

The effects of performance assessments on instructional practices are more complex. One of the rationales for the introduction of performance assessment in state testing programs was to signal instructional direction without narrowing the curriculum as multiple choice tests had done (Resnick & Resnick, 1992). There is some evidence that positive curriculum change has occurred. For example, portfolio assessment has lead to positive changes in teaching practices (Stecher & Herman, 1997). Principals and teachers in Maryland and Kentucky generally believe that test-based accountability has at least a small positive impact of instruction, and that accountability has caused them to focus on content and skills that are assessed (Koretz, Mitchell, Barron, & Keith, 1996; Koretz, Barron,

Mitchell, & Stecher, 1996). Similarly, teachers in Kentucky report increased focus on tested subjects and increased use of practices encouraged by the test reformers (Stecher, Barron, Kaganoff, & Goodwin, 1998). To the extent that the goal of test-based accountability is to focus on previously neglected content or skills (e.g., writing, problem solving), it appears to be generally successful.

The use of performance assessment, however, has not eliminated all of the negative instructional consequences associated with high stakes testing (Koretz, Stecher, Klein, & McCaffrey, 1994). For example, teachers in Vermont focused on the portfolio scoring rubrics rather than the domains of mathematics the assessment was supposed to measure (Stecher & Mitchell, 1995). Similarly, teachers in Maryland and Kentucky appear to focus inappropriately on test preparation activities that do not generalize to the curriculum as a whole (Koretz, Mitchell, Barron, & Keith, 1996; Koretz, Barron, Mitchell, & Stecher, 1996; Koretz & Barron, 1998). There also is evidence that high stakes performance assessments create conflicting pressures on teachers, who have a difficult time balancing the need to produce high scores with the desire to incorporate more authentic activities into their lessons (Borko & Elliott, 1999; Wolf & McIver, 1999).

Test-Based Accountability

Recently, a number of states have adopted formal school accountability systems that rely heavily on test scores. The popularity of formal test-based accountability is growing because it is seen by many as a relatively quick, relatively inexpensive, and highly visible way to bring about changes in schools. Policymakers hope such systems will encourage educational improvement by sending strong, clear signals to schools about their success.

However, the research on testing effects suggests that the manner in which an assessment system is structured will affect its utility as a tool for program improvement (Linn, 1999). For example, choices about high or low stakes, multiple choice, or performance assessment will have consequences in terms of teachers' behaviors and students' scores. In the accountability context, we would expect to find, further, that choices about grade levels and subject matter will make a difference, as will the nature of the adopted standards and the rewards or sanctions associated with performance. The present study examines changes in classroom practices in a state with a sophisticated test-based accountability system

that measures performance in selected subjects in selected "milepost" grades chosen from each school level (elementary, middle, and high school).

Before describing the study, we want to clarify what is meant by an accountability system. In simple terms, accountability is a relationship between two parties in which one party is expected to accomplish a particular goal and the other party is expected to provide benefits when the goal is accomplished (Hill & Bonan, 1991). A state accountability system is somewhat more complex because more parties are involved, goals are more abstract, and the flow of information is formalized.

Figure 1 is a model of a test-based state accountability system. The system involves relationships among four parties: state policymakers, school personnel, students and the public. State policymakers establish goals or expectations for students. Increasingly these goals take the form of performance standards that are

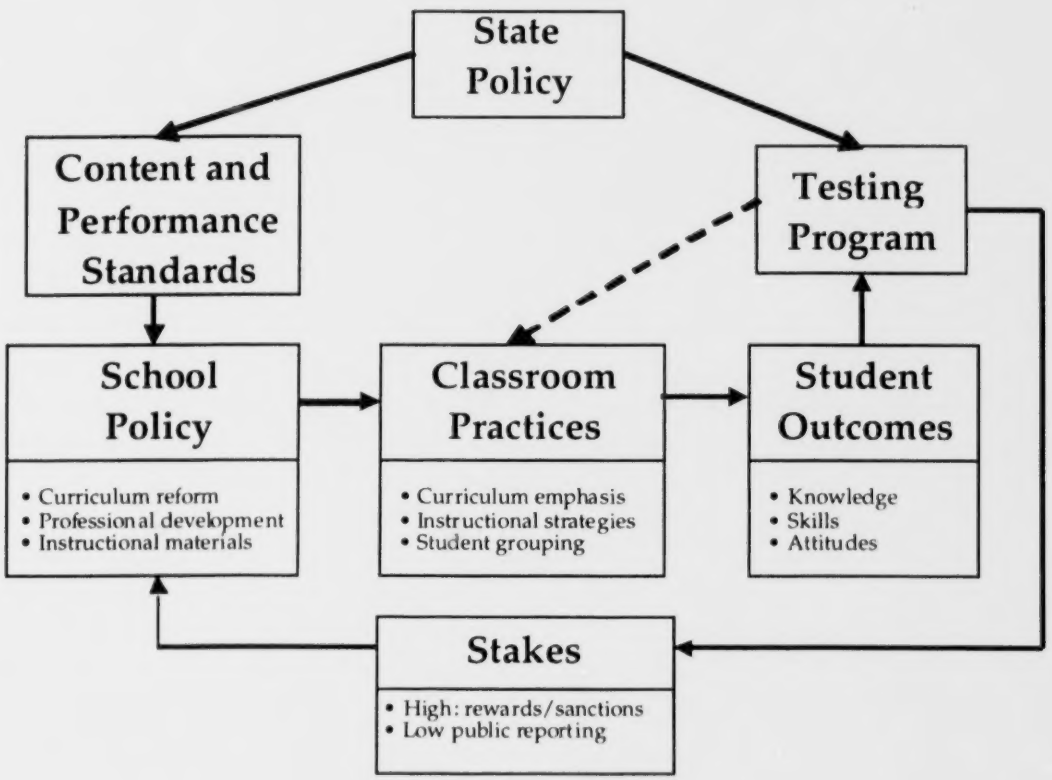


Figure 1. Test-based State Accountability System

adopted by state Boards of Education (Association of California School Administrators, 1996). It is important to note that most state standards are written at a high level of generality (e.g., communicate mathematics concepts effectively) which do not offer specific guidance about the content of lessons or the methods of instruction.

Schools provide the educational services that help students achieve the desired goals. School administrators set local policies and teachers implement specific classroom practices to promote student achievement. As a result of their classroom experiences, students acquire knowledge, master skills and develop attitudes toward learning. These student outcomes are compared to the standards to determine whether schools have been successful. Information about school performance is reported to the schools and to the general public. Schools enact changes based on these reports to improve the services they provide and enhance student outcomes. In addition, parents and community members may informally endorse or criticize the schools on the basis of the public information. In some systems, the state adds formal consequences. Successful schools are rewarded through recognition or financial incentives; unsuccessful schools are given assistance to help them improve. Under extreme circumstances, chronically unsuccessful schools may be reconstituted.

Tests play a critical role in this system. Students' normal school output cannot be easily translated into the language of the state standards. It is not possible to tell from grades, homework, and classroom work products whether students have achieved the desired performance standards. As a result, states create testing systems to measure student performance in ways that can be more easily judged against the standards. The school's accomplishments are measured and reported publicly using test results as indirect indicators of student accomplishment. The test defines the specific aspects of student performance that will be measured and reported.

However, practical constraints limit the amount and extent of testing that can be conducted. Consequently, test-based accountability systems actually measure a limited number of domains using a limited amount of data. In general, the selection of tests represents a compromise between practical considerations, political considerations, and the broad goals of the system.

Test results play a key role in the schools' responses to the accountability information, as well. The staff responds to published test results, rewards or sanctions, and other feedback by reinforcing practices perceived to be successful and modifying practices perceived to be unsuccessful. Ideally, the feedback will exert pressure on schools and teachers to make changes that will help students master the standards. Staff will focus their actions on promoting the broad goals endorsed by the state and student performance will improve. Evidence cited above shows that teachers do change their behaviors in response to such feedback.

Given limited time and resources, however, schools often direct their attention more narrowly to practices that will enhance student performance on the tests. This is one way in which the discrepancy between broad goals and specific measures may reduce the effectiveness of a test-based accountability system.

Of particular concern in this study is the use of a system that only tests certain subjects and only tests in selected, milepost grade levels. To minimize costs and testing burden, many states follow the example of NAEP and test at one grade level each in elementary, middle, and high school. This pattern lowers the testing burden on schools and the expense of the testing program. (Concerns about cost and testing burden are heightened with tests that contain open-response questions, which are more time-consuming to administer and to score.) Policymakers assume that local schools will translate information in the milepost grades into proper guidance for other grade levels. For example, the school would work backwards from fourth-grade objectives (if that is the accountability grade) to develop precursor skills and objectives in third, second, and first grade. They might also develop benchmarks for performance at the earlier grade levels. However, a more limited response would be one that focused narrowly on certain grades, subjects, topics within subjects, and achievement standards, but not the full range.

Test-Based Accountability in Kentucky

Kentucky provides an opportunity to study a high-stakes test-based accountability system that uses milepost testing to see how schools react to the accountability pressures. Kentucky has been in the forefront of the test-based accountability movement, and many states have looked to Kentucky when

designing their own accountability systems. Hence, lessons learned in Kentucky will be immediately relevant to other states.

The accountability model in Figure 1 fits the Kentucky system quite well, and we will briefly review each of the components. The Kentucky Educational Reform Act of 1990 established six broad goals for Kentucky schools. For example, the goal that relates to academic achievement states, "Schools shall develop their students abilities to use basic communication and mathematics skills for purposes and situations they will encounter throughout their lives" (Kentucky Department of Education, 1994, p. 2). A task force formed by the Kentucky Department of Education elaborated these into more detailed Academic Expectations that described what students should be able to do at the conclusion of their education. For example, "1.11: Students write using appropriate forms, conventions and styles to communicate ideas and information to different audiences for different purposes" (Kentucky Department of Education, 1994, p. 2). Later, KDE added further clarification in a document called *Transformations: Kentucky's Curriculum Framework* (Kentucky Department of Education, 1995). This included general descriptors of what would be appropriate at the elementary, middle school and secondary levels. In the case of writing, there are six elementary demonstrators, ranging from "Express thought/ideas through verbal and/or symbolic representation (e.g., pictures, scribbles, words)" to "Establish and use criteria for effective writing to evaluate own and others' writing" (Kentucky Department of Education, 1995, p. 26). Further clarification was issued in 1996 in a document called Core Content for Assessment (Kentucky Department of Education, 1996), but these remain general descriptions.

Kentucky developed their own assessment system to measure progress toward meeting its Academic Expectations. The Kentucky assessment system is perhaps the most elaborate in the country.[2] At the time this study was conducted, seven subjects were tested, and each was measured in one elementary grade level, one middle school grade, and one high school grade. The Kentucky assessment emphasized performance assessments, including portfolios and open-response measures. To reduce pressure on individual students and

[2] Until recently, the Kentucky assessment system was known as the Kentucky Instructional Results Information System (KIRIS), and many people are familiar with this acronym. In 1998 the system was reformed and the name was changed to the Commonwealth Achievement Testing System (CATS).

emphasize improvement as a whole-school activity, the system only reported data at the school level. The school accountability index also included non-cognitive measures (including attendance, drop-out rates, etc.) although they contributed just one-sixth of the total score and showed very little variability.

The Kentucky system has undergone many changes over the years. Originally all subjects were assessed in three grades, but the burden of portfolios and constructed response items was so great some of the testing shifted to adjacent grades. In 1996-97, four subjects were tested in grades 4, 7, and 11. Three others were tested in grades 5, 8, and 11. More recently multiple choice tests with individual scores have been added.

The Kentucky accountability system provides both informational feedback and consequences. In fact, Kentucky is a good example of an accountability system with high stakes for schools. Scores are published and widely disseminated. Schools can receive financial rewards to be distributed among the staff if student scores on the state assessment exceed improvement targets, and the schools are subject to review and external intervention if students' performance is consistently poor. The level of performance needed to receive rewards is tied to continual improvement not absolute attainment. This approach puts pressure on all schools because even those who are scoring the highest have to show improvement in the next accountability cycle.

Finally, Kentucky provides extensive professional development opportunities to help schools and teachers improve student performance. A network of regional centers was created to help teacher understand the academic expectations and the new assessment system and to integrate them into instruction.

Kentucky has been implementing test-based accountability for almost a decade, making it a good site for studying the effects of the milepost testing model. Teachers have had ample opportunity to become familiar with the Academic Expectations and the format in which they are measured. Kentucky provides a unique opportunity to see the degree to which accountability reactions generalize beyond the tested grade level, and to see whether teachers respond to the narrow signals of specific tests or the larger targets embodied in the academic standards.

Procedures

In 1996, RAND undertook a two-year study of the impact of standards-based assessment on classroom practices in Kentucky. Kentucky teachers were surveyed on their classroom practices as well on other school practices during the 1996-97 school year and during the 1997-98 school year. Also during this time period, case studies of a small group of exemplary teachers were conducted. This paper is based on the survey results of the second year of that effort. Results of the case studies and the first year survey are presented elsewhere (Borko & Elliott, 1999; Wolf & McIver, 1999; Stecher, Barron, Kaganoff, & Goodwin, 1998).

The 1997-98 RAND/CRESST survey of Kentucky teachers involved writing and math teachers from grades 4-7. We selected these subjects because of their importance in the Kentucky education reform and because both were assessed using portfolios, one of the more innovative components of KIRIS. Grades 4-7 were selected in order to obtain responses from teachers in accountability grades (grades 4 and 7 for writing, grade 5 for math) as well as in a non-accountability grade (grade 6). This report summarizes the results of the survey that pertain to differences in practice related to the accountability burden in each grade.

Sampling

Kentucky schools were classified into two overlapping groups based on the grade levels that were taught in the school. All schools containing grades 4 and 5 were included in the elementary school sample and all schools containing grade 6 and 7 where included in the middle school sample (some schools; e.g., K-8 schools) were included in both groups whereas others (e.g., K-4) were not included in either group. We were interested in obtaining information from teachers in both accountability and non-accountability grades about the pressure they feel to prepare their students for the KIRIS tests the students will be taking in the following grade. Grade 4 is a not an accountability grade for math but grade 5 is so we wanted to sample fourth-grade teachers in schools that also contained fifth grade. Similarly, sixth grade not an accountability grade in writing but seventh grade is so we wanted to sample sixth-grade teachers in schools that also contained seventh grade.

For each population, schools were divided into four strata of equal size based on average enrollment in the grades of interest. Schools with fewer than 20 students in the accountability grade were excluded from the sampling frame, as

were schools with recent changes in their service areas. Within each stratum a random sample of schools was chosen. Seventy-two schools were selected for the elementary school sample and eighty schools were selected for the middle schools sample. No school was chosen for more than one sample.

A letter was sent to the principal of each school at the beginning of 1998 explaining the study and requesting the names of the instructors teaching the identified grades and subjects. Principals were subsequently contacted by telephone to retrieve these names. Ninety-three percent of the principals in the sampled schools provided the requested information.

The teachers were contacted by mail and asked to participate in the study. The contact letter explained the study and asked for their participation. Enclosed with the request were a letter from the Department of Education urging teacher cooperation, a copy of the survey to be completed, a return envelope, and a pre-paid $10 long distance phone card. Teachers could keep the phone card whether or not they returned the survey.

Four hundred and seventy-nine teachers completed the survey for an overall response rate of 54 percent. This was lower than the response rate achieved in previous RAND surveys of Kentucky teachers. Several explanations for the lower response rate are plausible. First, the survey was mailed to teachers near KIRIS testing time so some teacher may have felt they just didn't have the time to complete it. Second, there has been considerable research conducted in Kentucky and some teachers may have grown weary of survey requests. Third, when the survey was in the field, the state legislature was deciding to eliminate the KIRIS program and adopt a new accountability program. Thus teachers may have felt that with the KIRIS system on the way out, the survey results would be of little use.

Survey Design and Data Collection

Building on past RAND research, the surveys addressed a broad range of issues related to classroom practices. Major themes included professional development, school and class organization, curriculum and instruction, test preparation, and school level practices related to the accountability assessments.

Most of the survey questions were presented in a closed format. Respondents were asked to provide numerical answers or to select one option from a predetermined set of options (e.g., three-, four-, and five-point Likert

scales, and yes/no questions). We also asked a number of open-response questions that allowed respondents to explain or expand upon their answers to the closed-ended questions.

Change was a predominant theme in the survey. For most questions about practice, teachers were asked about current behaviors (during the 1997-98 school year) and about changes during the past three-year period. Only teachers with at least three years of experience answered questions about changes in practice. Twelve percent of the elementary teachers and 18 percent of the middle school teachers indicated that they could not answer these questions.

Analysis

Because we were interested in comparing teachers classroom practices across grades, only teachers who teach a single grade were included in the analyses reported here (N=365).

All analyses were conducted using weighted data. The weight assigned to each case was the product of the inverses of the probability that the school would be selected and probability that the sampled individuals would participate (complete the survey). Descriptive statistics were calculated overall and separately for each grade. When data were combined across grades, the grades were weighted equally in the combined statistics. For the Likert questions, frequencies were computed. For questions requiring a numerical response, we calculated means, medians, and standard deviations.

We tested the significance of the differences between responses for teachers in different grades. The majority of statistical tests performed were chi-squared tests comparing responses across grades where the responses were dichotomized (e.g., no/yes, low/high, or less frequently/more frequently). For questions requiring a numerical response one-way analyses of variance (ANOVA) were used to compare groups.

There were a number of open-response questions on the survey. Responses to these questions were read by project staff, and codes were developed for all responses that occurred with any regularity. The responses were then coded and the codes tallied. Responses to these questions are only used for descriptive purposes – no significance tests were carried out.

Results

There were strong associations between the grade levels at which specific subjects were assessed in KIRIS and the educational practices of teachers in those grades. Differences related to KIRIS grade levels were found in teachers' participation in professional development, their allocation of instructional time across subjects (in self-contained classrooms), and the relative emphasis they placed on specific topics within the subjects of mathematics and writing.

Table 1 shows the KIRIS testing grades by subject for 1997-98. These are the milepost subject-by-grade combinations that serve as the reference for comparisons among the survey results for teachers in different grades. Mathematics portfolios have been part of the KIRIS assessment system off and on. During the 1997-98 school year, they were not officially being collected or scored. However, many teachers did have their students compose mathematics portfolios in anticipation of their return to the accountability system in future years.

Table 1

Assessment Grade Levels for KIRIS, 1997-98

Subject	Grade			
	Fourth	Fifth	Sixth	Seventh
Reading	O			O
Writing	P			P
Science	O			O
Mathematics		OP*		
Social Studies		O		
Arts and Humanities		O		
Practical living/Vocational education		O		

Note. P designates cumulative portfolio assessments; O designates on-demand, open-response assessments.

*KIRIS mathematics portfolios were not in use at fifth grade in 1997-98. However, because they were scheduled to be added to the KIRIS mathematics assessment at fifth grade in the future, many teachers were using them at least informally.

Professional Development

Teachers in all grades reported participating in 50 to 70 hours of professional development during the past year. However, sixth-grade teachers participated in fewer hours of professional development on average (57.6) than did teachers in the KIRIS accountability grades of 4, 5, and 7 (69.1, 64.4, and 65.4, respectively), although this difference was not large enough to be statistically significant.[3] Most teachers in all four grades reported that their professional development emphasized assessment, particularly the KIRIS assessments (see Table 2). The assessment topics reported by the greatest percentage of teachers were: preparation for KIRIS portfolios (81 percent); preparation for KIRIS open-response items (76 percent); and scoring KIRIS portfolios (70 percent). The only significant differences between grade levels in professional development were related to portfolio scoring. Teachers in grade 4 reported spending more professional development time on the scoring of portfolios than teachers in other grade levels. And even at grades where there was no official portfolio scoring (grades 5 and 6), more than half of the teachers reported receiving more than a small amount of training on portfolio scoring.

Table 2

Percent of Teachers Who Reported Spending a Moderate Amount or Great Deal of their Professional Development on Each Area

	Grade			
	Fourth	**Fifth**	**Sixth**	**Seventh**
Preparation for KIRIS open response	**65**	**73**	82(4)	**82**
Preparation for KIRIS multiple choice	**14**	**23**	22	**14**
Preparation for KIRIS portfolios	**89**	**77**	75	**83**
Scoring of KIRIS portfolios	**86**(567)	**61**(4)	68(4)	**66**(4)
Classroom assessment methods	**31**	**44**	45	**41**
Use of education technology	**43**	**54**	47	**36**

Note. Accountability grades are in bold.
(4) Significantly different from Grade 4 (α= .01).
(5) Significantly different from Grade 5 (α = .01).
(6) Significantly different from Grade 6 (α = .01).
(7) Significantly different from Grade 7 (α = .01).

[3] Due to the skewness of the responses, differences between grades were tested using a log transformation.

Elementary teachers in grades four and five who teach all subjects received more professional development on the subjects that were assessed in their grade level than on the subjects that were not assessed in their grade (see Table 3).[4] For example, almost all fourth-grade teachers reported attending professional development that emphasized writing and one-half attended professional development that emphasized science compared to one-third or fewer who attended professional development that emphasized mathematics or social studies. For the most part the opposite was true for fifth-grade teachers. They were more likely to participate in professional development that emphasized mathematics and social studies than science. The differences between the grade levels were statistically significant for writing, science, and mathematics.

Table 3

Percent of Teachers in Self-Contained Classrooms Who Reported Spending a Moderate Amount or Great Deal of their Professional Development on Each Subject

	Grade	
	Fourth	Fifth
Subjects tested in fourth grade		
Reading	38	42
Writing*	96	84
Science*	50	28
Subjects tested in fifth grade		
Mathematics**	35	87
Social Studies	27	41

* Significant at α =.05.
**Significant at α =.01.

There were exceptions to pattern of professional development aligning with assessment grade level. Over 80 percent of fifth-grade teachers reported attending professional development that emphasized writing even though writing, per se, is not tested in fifth grade. This interest in writing may be explained, in part, because the KIRIS assessments in other subject areas require a great deal of writing. In fact, the mathematics portfolios demand a considerable amount of

[4] All analyses were restricted to teachers in fourth and fifth grades who teach all academic subjects. Data from sixth and seventh grade teachers were not used because so few teachers at these grades taught all subjects.

writing of a specialized nature. Even so, the percentage of fifth-grade teachers who reported their professional development emphasized writing was still significantly lower than the percentage of fourth-grade teachers. Less easy to explain is the fact that only 38 percent of fourth-grade teachers reported that their professional development emphasized reading compare to 42 percent of fifth-grade teachers.

Allocation of Time by Subject in Self-Contained Classrooms

Teachers who were responsible for teaching all subjects allocated classroom time to subjects in ways that reflected the KIRIS accountability milepost grades. Fourth- and fifth-grade teachers in self-contained classrooms reported allocating more time to subject areas that were tested in their grade levels than to other subjects. They also reported increasing the amount of time allocated to tested subjects while decreasing the amount of time allocated to other non-tested subjects.

The results in Table 4 show consistent differences between fourth and fifth grade in the amount of instructional time devoted to each subject, and these differences mirror the subjects tested by KIRIS. In all subjects except reading, teachers in the accountability grade for a subject spend significantly more hours per week on that subject than teachers in the non-accountability grade. For

Table 4

Mean Time Spent on Each Subject in a Typical Week Reported by Teachers in Self-Contained Classrooms

	Grade	
	Fourth	Fifth
Subjects tested in fourth grade		
Reading	5.2	4.7
Writing**	5.8	4.0
Science**	5.2	3.5
Subjects tested in fifth grade		
Mathematics**	4.9	6.4
Social Studies**	3.5	5.6
Arts and Humanities**	1.5	2.4
Practical living/Voc. educ.**	1.4	2.4

*Significant at α =.05.
**Significant at α =.01.

example, teachers of fifth grade reported spending 60 percent more time (5.6 hours per week vs. 3.5 hours per week) on social studies (which is tested in fifth grade) than teachers of fourth grade. Conversely, science, which is tested by KIRIS in fourth grade, received 49 percent more classroom time from fourth-grade teachers (5.2 hours per week) than from fifth-grade teachers (3.5 hours per week).

Teachers reported similar differences when asked about changes in the allocation of instructional time across subjects. Overall, more teachers reported increasing the amount of time spent on subjects that were tested at their grade level than subjects that were not tested (see Table 5). In all cases except reading and arts and humanities, the percentage of teachers in the accountability grade who reported increasing the time spent on each subject was significantly greater than the percentage of teachers in the non-accountability grade. For example, 59 percent of teachers of fifth grade reported increasing the amount of time they spent on social studies whereas only 18 percent of teachers of fourth grade reported increasing the time spent on social studies.

Table 5

Percent of Teachers in Self-Contained Classrooms Who Reported Increasing the Time Allocated to Each Subject Area

	Grade	
	Fourth	Fifth
Subjects tested in fourth grade		
Reading	23.3	31.7
Writing*	80.2	61.7
Science**	76.3	13.1
Subjects tested in fifth grade		
Mathematics**	13.6	81.9
Social Studies**	18.0	59.1
Arts and Humanities	42.1	54.8
Practical living/Voc. educ.*	29.8	54.5

* Significant at α =.05.
**Significant at α =.01.

In response to an open-ended question, teachers confirmed that KIRIS was the primary reason for the changes they made in the use of classroom time. In fourth grade, 84 percent of the teachers who responded to this question indicated

that the KIRIS assessment were responsible for the changes they made. Most of these teachers (62 percent) specifically mentioned the KIRIS writing portfolios. Typical of these responses is the following, "Since fourth grade is responsible for portfolios at the elementary level, my time spent on writing has increased tremendously since last year. Since writing time has increased, I've had to slack off on other subjects, especially since Christmas." In addition, 38 percent of the teachers who attributed changes to KIRIS mentioned the trade-off between tested and non-tested subjects. For example, one teacher wrote, "[We] spend a great deal of time teaching content areas covered on the KIRIS test, other academic areas suffer as a result." Another teacher explained, "only reading, writing, and science are on the fourth-grade test."

Similarly, 86 percent of the fifth-grade teachers who responded to the open-ended question ascribed the changes in instructional time to KIRIS. However, unlike fourth grade where the writing portfolios were very influential, relatively few of the fifth-grade teachers (17 percent) ascribed changes to the mathematics portfolios. Explanations from fifth-grade teachers attribute the change to the subjects tested at the fifth grade (50 percent) or to KIRIS in general without going into detail (33 percent). Typical of fifth-grade responses are the following: "changes in instructional time devoted to any subject areas were a result of the pressures to prepare students for KIRIS assessments"; and "I have made a more conscious effort to include aspects of arts and humanities and [sic] practical living and vocational studies (especially since these are tested in fifth grade)."

Mathematics Curriculum and Instruction

There were differences in the curriculum and instructional practices of teachers who taught mathematics depending on whether they taught in an accountability grade or a non-accountability grade. In particular, the responses of fifth-grade teachers were significantly different from the responses of teachers in the other grades to questions about mathematics curriculum coverage, instructional strategies, student learning activities, and KIRIS preparation activities.

All teachers who taught mathematics in grades 4, 5, 6, or 7 reported on the frequency with which they covered four core content areas of mathematics: numbers and computation, geometry and measurement, statistics and probability, and algebraic ideas (see Table 6). Significantly more fifth-grade

Table 6

Percent of Teachers Who Reported Covering Each Mathematics Content Area at Least Once a Week

	Grade			
	Fourth	**Fifth**	Sixth	Seventh
Numbers and computation	98	**97**	100	94
Geometry and measurement	26[5]	**56**[46]	29[5]	32
Statistics and probability	22[56]	**50**[467]	4[45]	18[5]
Algebraic ideas	31[5]	**56**[46]	27[57]	56[6]

Note. Mathematics accountability grade in bold.
[4] Significantly different from grade 4 (α = .01).
[5] Significantly different from grade 5 (α = .01).
[6] Significantly different from grade 6 (α = .01).
[7] Significantly different from grade 7 (α = .01).

teachers reported covering statistics and probability on at least a weekly basis than did teachers in the non-accountability grades. More fifth-grade teachers reported covering geometry and measurement on at least a weekly basis than did teachers in the non-accountability grades—the differences were significant at α=.01 for fourth and sixth grade and α=.02 for seventh grade. Also, more fifth-grade teachers reported covering algebraic ideas on at least a weekly basis than did teachers in fourth and sixth grade. There was very little difference between fifth- and seventh-grade teachers in terms of their coverage of algebraic ideas, which is probably explained by the fact that algebraic ideas are traditionally introduced to students in seventh grade. Almost all teachers in all grades reported covering numbers and computation at least weekly and the difference between the grades was not significant.

A similar pattern of increased coverage in the accountability grade was reported for elements of mathematics that are emphasized in current mathematics reforms, such as connections among mathematics ideas and mathematical communication (see Table 7). The accountability effect was seen most clearly with regards to mathematics communication where teachers in fifth grade reported greater emphasis than teachers in the non-accountability grades. Teachers in fifth grade reported more emphasis on connections among mathematics ideas and connections between mathematics and other subjects, however, the effects were not always large enough to be significant at α=.01.

Table 7

Percent of Teachers Who Reported Covering Each Mathematics Area at Least Once a Week (Accountability Grade in Bold)

	Grade			
	Fourth	**Fifth**	Sixth	Seventh
Connections among mathematics ideas	81	**93**[7]	78	74[5]
Connections between mathematics and other subjects	74	**84**[67]	53[5]	61[5]
Problem solving	97[7]	**96**[7]	87	79[45]
Mathematical communication	62[5]	**96**[467]	73[5]	76[5]
Reasoning	94	**93**	90	77
Technology	72[6]	**74**[6]	44[45]	58

Note. Mathematics accountability grade in bold.
[4] Significantly different from Grade 4 ($\alpha = .01$).
[5] Significantly different from Grade 5 ($\alpha = .01$).
[6] Significantly different from Grade 6 ($\alpha = .01$).
[7] Significantly different from Grade 7 ($\alpha = .01$).

The accountability-related differences between grade levels extended to some instructional strategies, as well. Teachers in the accountability grade (grade 5) were more likely to engage in selected instructional activities frequently (at least once a week) than teachers in other grades, although this was not true for all the activities we examined (see Table 8). A significantly greater percentage of teachers in fifth grade reported asking open-response questions with many right answers than did teachers in grades 4, 6, or 7. Although most teachers in all grades reported frequently demonstrating how to perform a new mathematics skill, more teachers in fifth grade reported doing this activity frequently than did teachers in the non-accountability grades. Also, more teachers in fifth grade reported demonstrating mathematics ideas using constructions, manipulatives, etc., and showing connections between mathematics and other subjects than did teachers in sixth and seventh grades where as the difference between grade 4 and grade 5 was not large enough to be significant at α=.01.

Table 8

Percent of Teachers Who Reported Doing Each Activity Frequently (at Least Once a Week)

	Grade			
	Fourth	**Fifth**	Sixth	Seventh
Demonstrate how to perform a new mathematics skill	88[5]	**100**[467]	90[5]	87[5]
Ask open-response questions with many right answers	47[5]	**77**[467]	41[5]	23[5]
Explain a new concept	81	**90**	91	80
Give examples of real-world applications of mathematics skills	93	**91**	81	84
Demonstrate mathematics ideas using constructions, manipulatives, etc.	71[6]	**87**[67]	43[45]	56[5]
Conduct speed drills	41[567]	**15**[4]	8[4]	6[4]
Explain correct solutions to assigned problems	98	**100**	100	98
Give tests or quizzes	40	**49**	44	54
Show connections between mathematics and other subjects	80[6]	**82**[67]	58[45]	60[5

Note. Mathematics accountability grade in bold.
[4] Significantly different from Grade 4 ($\alpha = .01$).
[5] Significantly different from Grade 5 ($\alpha = .01$).
[6] Significantly different from Grade 6 ($\alpha = .01$).
[7] Significantly different from Grade 7 ($\alpha = .01$).

Differences between grade levels in the time allocated to mathematics, the curriculum emphasized in mathematics, and the mathematics instructional behaviors teachers engaged in were mirrored in differences in student learning activities (see Table 9). Fifth-grade teachers reported having students engage in almost all of the activities more frequently than teachers in non-accountability grades. On average, seven percent more teachers in accountability grades than non-accountability grades reported having students engage in the activity at least once a week. However, there were several types of activities in which fifth-grade teachers reported considerably higher frequencies than did teachers in non-accountability grades. The largest differences concerned having students write about mathematics and having them represent concepts or ideas in tables, graphs, or pictures. A majority of fifth-grade teachers had students engage in these activities at least weekly, whereas far fewer teachers in non-accountability grades had their students engage in these activities at least weekly. Similar results, but not as dramatic, were found for other activities, including: (a) work

on extended mathematics activities that take several days; (b) use mathematics in the context of other subjects; (c) discover mathematics concepts for themselves; and (d) use measuring tools.

Table 9

Percent of Teachers Who Reported Having Students Engage in Each Activity Frequently (at Least Once a Week)

	Grade			
	Fourth	**Fifth**	Sixth	Seventh
Practice computation skills	97[57]	**84**[4]	92[7]	70[46]
Write about mathematics	33[5]	**67**[46]	32[5]	47
Represent concepts or ideas in tables, graphs, or pictures	42[5]	**68**[467]	33[5]	36[5]
Solve problems using manipulatives	54[6]	**59**[6]	33[45]	40
Use mathematics to solve real-world problems	77	**89**	75	69
Work on extended mathematics activities that take several days	5[5]	**32**[47]	18	6[5]
Learn mathematics facts, rules, or formulas	69	**78**	68	59
Use mathematics in the context of other subjects	47	**65**[67]	40[5]	34[5]
Discover mathematics concepts for themselves	36	**56**[67]	22[5]	26[5]
Work problems from the textbook	89	**85**	79	76
Use measuring tools in mathematics	54[6]	**63**[67]	29[45]	37[5]
Explain their thinking to other students	70	**76**	63	72
Work on problems in groups with other students	52	**64**	47	51

Note. Mathematics accountability grade in bold.
[4] Significantly different from Grade 4 ($\alpha = .01$).
[5] Significantly different from Grade 5 ($\alpha = .01$).
[6] Significantly different from Grade 6 ($\alpha = .01$).
[7] Significantly different from Grade 7 ($\alpha = .01$).

There were also large differences between fifth-grade teachers and teachers in the non-accountability grades in terms of activities aimed specifically at improving KIRIS performance (see Table 10). Fifth-grade teachers were more likely than were teachers in other grades: to have students frequently practice using KIRIS released items; show students responses for KIRIS or KIRIS-like items that illustrate different levels of performance; to have students assign proficiency levels to classroom work; to practice scoring portfolio pieces; and to frequently use open-response items as part of their class work. Interestingly, teachers in all grades were likely to display visual aids in their classroom for

students to refer to when working on KIRIS-like tasks—overall, 77 percent of teachers reported that visual aids are displayed at least weekly.[5]

Table 10

Percent of Teachers Who Reported Engaging in Each Activity Frequently (at Least Once a Week) to Help Students Do Well on KIRIS Mathematics Assessment

	Grade			
	Fourth	**Fifth**	Sixth	Seventh
Have students practice using KIRIS released items	36(67)	**54**(67)	14(45)	8(45)
Show students responses from KIRIS items or KIRIS-like items that illustrate different levels of performance and discuss them	22(5)	**47**(467)	11(5)	12(5)
Have students assign proficiency levels to classroom work	19(5)	**41**(467)	8(5)	10(5)
Practice scoring portfolio pieces	9	**20**(67)	0(5)	2(5)
Use open-response items in classwork	45(5)	**67**(467)	31(5)	22(5)
Display scoring rubrics in the classroom	63(67)	**54**(6)	27(45)	31(4)
Display visual aids in your classroom for students to refer to when working on KIRIS-like tasks	71	**82**	74	85

Note. Mathematics accountability grade in bold.
(4) Significantly different from Grade 4 ($\alpha = .01$).
(5) Significantly different from Grade 5 ($\alpha = .01$).
(6) Significantly different from Grade 6 ($\alpha = .01$).
(7) Significantly different from Grade 7 ($\alpha = .01$).

Mathematics teachers did not respond in large numbers to an open-ended question about changes in their mathematics curriculum. Of those who did respond, the largest number of teachers (19 percent) indicated that the biggest change they made was to increase their emphasis in problem solving and reasoning. Other changes mentioned by a number of teachers were more writing (15 percent), more open-response questions (13 percent), more manipulatives (13 percent). Sixty-five percent of the teachers who described a change in curriculum or instruction said the change was motivated by KERA or KIRIS. This was true in all grades not just in the mathematics accountability grade.

[5] It is possible that teachers interpreted this question to refer to KIRIS-like tasks in subjects other than mathematics.

Writing Curriculum and Instruction

As was the case with mathematics, the content of writing instruction and the instructional activities teachers use differ between accountability and non-accountability grades.

There are significant differences in the emphasis teachers place on various aspects of writing (see Table 11). More teachers in the accountability grades

Table 11

Percent of Teachers Who Reported Covering Each Content Area at Least Once a Week

	Grade			
	Fourth	**Fifth**	**Sixth**	**Seventh**
Awareness of audience	**86**(56)	50(47)	60(47)	**87**(56)
Focused purpose	**89**(56)	64(47)	66(47)	**92**(56)
Tone/voice	**83**(56)	52(4)	47(4)	**67**
Idea development; use of supporting details	**96**(56)	68(47)	76(47)	**95**(56)
Logical organization; use of transitions	**94**(56)	66(4)	70(4)	**84**
Sentence structure	**97**(567)	79(4)	83(4)	**81**(4)
Use of effective language	**92**	84	81	**81**
Mechanics	**97**(567)	80(4)	81(4)	**80**(4)
Writing in a variety of genres/forms	**69**	59	44	**67**

Note. Writing accountability grades in bold.
(4) Significantly different from Grade 4 ($\alpha = .01$).
(5) Significantly different from Grade 5 ($\alpha = .01$).
(6) Significantly different from Grade 6 ($\alpha = .01$).
(7) Significantly different from Grade 7 ($\alpha = .01$).

(grades 4 and 7) reported covering awareness of audience, focused purpose, and idea development than teachers in the non-accountability grades (grades 5 and 6). Similar trends were observed for tone/voice and logical organization, however, the differences between grades 5 and 6 and grade 7 were not large enough to be significant at α=.01. Sentence structure and mechanics were covered frequently by more fourth-grade teachers than teachers in the other grades including grade 7, possibly indicating that by grade 7, teachers feel that students have mastered these basic skills.

Although higher frequencies in the accountability grades than in non-accountability grades were observed for a number of the instructional activities we asked about, typically the differences were only significant (α=.01) between

fourth grade and the non-accountability grades (see Table 12). Teachers in the lower accountability grade (grade 4) were more likely to report that they frequently demonstrate the use of pre-writing and give examples of choosing appropriate words to describe objects or experiences. Teachers in grade 4 were more likely than teachers in all other grades to provide time for unstructured student writing, suggest revisions to student writing, and to show examples of writing in different content areas on a frequent basis.

Table 12

Percent of Teachers Who Reported Frequently (at Least Once a Week) Doing Each Activity

	Grade			
	Fourth	**Fifth**	**Sixth**	**Seventh**
Read orally to students	**90**[67]	94[67]	72[45]	**64**[45]
Use examples to discuss the craft of an author's writing	**76**[6]	63	49[4]	**70**
Provide time for unstructured student writing	**87**[567]	55[4]	64[4]	**68**
Demonstrate the use of pre-writing (e.g., webbing)	**76**[56]	54[4]	43[4]	**56**
Suggest revisions to student writing	**96**[567]	61[4]	60[4]	**74**[4]
Give examples of choosing appropriate words to describe objects or experiences	**89**[56]	62[4]	67[4]	**73**
Explain correct usage of grammar, spelling, punctuation, and syntax	**95**	87	86	**90**
Give tests or quizzes	**24**[6]	33	52[4]	**30**
Show examples of writing in different content areas	**72**[567]	49[4]	34[4]	**34**[4]
Write with students on same assignment	**34**	31	19	**25**

Note. Writing accountability grades in bold.
[4] Significantly different from Grade 4 ($\alpha = .01$).
[5] Significantly different from Grade 5 ($\alpha = .01$).
[6] Significantly different from Grade 6 ($\alpha = .01$).
[7] Significantly different from Grade 7 ($\alpha = .01$).

Writing teachers in accountability grades ask their students to produce longer written pieces—one or more pages—more often than writing teachers in non-accountability grades. As Table 13 shows, the vast majority of teachers in all grades have their students produce written pieces of one to two paragraphs in length at least once a week. However, teachers in writing accountability grades are more likely to have their students produce pieces of longer lengths—one to

two pages—at least once a week and pieces of three or more pages at least once a month.

Table 13

Percent of Teachers Who Reported Frequently (at Least Once a Week) Having Students Produce Written Pieces of Specified Length

	Grade			
	Fourth	Fifth	Sixth	**Seventh**
At least once a week				
One to two paragraphs	**91**	93	79	**80**
One to two pages	**68**[56]	36[4]	21[47]	**47**[6]
Three or more pages	7	9	7	**15**
At least once a month				
Three or more pages	**61**[56]	40[47]	34[47]	**74**[56]

Note. Writing accountability grades in bold.
[4] Significantly different from Grade 4 ($\alpha = .01$).
[5] Significantly different from Grade 5 ($\alpha = .01$).
[6] Significantly different from Grade 6 ($\alpha = .01$).
[7] Significantly different from Grade 7 ($\alpha = .01$).

The writing process approach is widely used to teach writing in all four grades, and student use most of the steps of the writing process on the majority of the pieces they write. However, there were grade-to-grade differences in the percentage of teachers who regularly use many of the writing process steps. Table 14 shows the percentage of teachers who indicated that students in their classroom engage in each writing process activity on more than half of the written pieces they produce. Although there was a tendency for a greater percentage of teachers in the accountability grades than the non-accountability grades to regularly use many of the steps, the comparisons that were significant tended to be between grade 5 and grades 4 and 7. This may reflect the fact that fifth-grade teachers focus on writing geared towards the KIRIS open-response tests and portfolios in mathematics whereas teachers in grades 4 and 7 tend to focus on writing geared towards the writing portfolios.

Table 14

Percent of Teachers Who Reported that Students Engage in Each Activity on More than Half of their Written Pieces

	Grade			
	Fourth	**Fifth**	**Sixth**	**Seventh**
Gather information/conduct research before they write	**10**	25	13	**10**
Pre-write (e.g., make a web, map, etc.)	**85**	72[7]	89	**95**[5]
Define the purpose and audience	**86**[57]	63[47]	79[7]	**100**[456]
Use conferencing with peers to improve their writing	**82**[5]	51[467]	76[5]	**87**[5]
Use conferencing with the teacher to improve their writing	**76**	61[7]	69	**83**[5]
Revise the piece at least once	**85**	73[7]	87	**91**[5]
Edit the piece to correct errors in mechanics	**86**	71[67]	91[5]	**95**[5]
Publish the piece for others to read	**71**	52	55	**66**

Note. Writing accountability grades in bold.
[4] Significantly different from Grade 4 (α = .01).
[5] Significantly different from Grade 5 (α = .01).
[6] Significantly different from Grade 6 (α = .01).
[7] Significantly different from Grade 7 (α = .01).

Unlike mathematics, we did not find large differences between teachers in the accountability and non-accountability grades in terms of activities designed to improve students' KIRIS performance in writing (Table 15). Instead, differences were more pronounced between teachers in elementary and middle schools. Elementary school teachers focused more on illustrating different levels of performance using responses to KIRIS or KIRIS-like items, teaching the four-column method, having students assign proficiency level to classroom work, and practicing scoring portfolio pieces. Teachers in all grades tended to display the scoring rubrics and other visual aids in their classrooms aimed at helping students with KIRIS-related writing.

Table 15

Percent of Teachers Who Reported Engaging in Each Activity Frequently (at Least Once a Week) to Help Students Do Well on KIRIS Writing Assessment

	Grade			
	Fourth	**Fifth**	**Sixth**	**Seventh**
Show students responses from KIRIS items or KIRIS-like items that illustrate different levels of performance and discuss them	**58**[67]	50[67]	**8**[45]	**22**[45]
Teach the four-column method	**48**[67]	45[7]	**24**[4]	**16**[45]
Have students assign proficiency levels to classroom work	**39**	43[7]	20	**19**[5]
Practice scoring portfolio pieces	**29**[7]	17	10	**7**[4]
Display scoring rubrics in the classroom	**76**[5]	56[4]	63	**77**
Display visual aids in your classroom for students to refer to when working on KIRIS-like tasks	**89**[5]	68[4]	88	**87**

Note. Writing accountability grades in bold.
[4] Significantly different from Grade 4 ($\alpha = .01$).
[5] Significantly different from Grade 5 ($\alpha = .01$).
[6] Significantly different from Grade 6 ($\alpha = .01$).
[7] Significantly different from Grade 7 ($\alpha = .01$).

Few writing teachers responded to our open-ended questions about changes in their writing program. The largest number of teachers (17 percent) responded that the biggest change they made was to increase their emphasis audience and purpose. Other changes mentioned by a large number of teachers were more types of writing—genres (13 percent), spending more time on writing (12 percent). Several teachers in the accountability grades reporting doing more portfolio work, which was not reported by teachers in the non-accountability grades. As in mathematics, when asked what prompted them to make changes to their curriculum, the majority of the writing teachers who responded (70 percent) gave an answer involving KERA or KIRIS. In addition, 15 percent of teachers who responded said they were prompted to make changes due to students' lack of skills or knowledge.

Discussion

This study confirms some of the positive effects of test-based accountability that have been reported previously (e.g., Stecher et al., 1998), but it also reveals previously unexamined negative consequences arising from high stakes

accountability systems using milepost testing. On the positive side, teachers react to KIRIS by changing their behaviors in ways that are consistent with the specific targets of the system. For example, Kentucky teachers participate in professional development activities that are consistent with the KIRIS-tested subjects. Furthermore, they emphasize the relevant content areas within the domains being assessed, and they plan lessons to improve students' abilities to demonstrate relevant skills. Those who teach in self-contained classes reallocate classroom time to emphasize the tested subjects.

On the negative side, teachers focus on the most proximal aspects of the system (tests) rather than the more distant goals it us supposed to promote (curriculum and performance standards). This "near-sightedness" manifests itself in many ways. Teachers attend primarily to the dimensions of the system that are relevant to them (e.g., the subjects measured at their grade level) rather than the system as a whole (all the subjects relevant to state standards). They focus on the measures that are reported (KIRIS items and portfolio pieces) rather than the constructs that are being measured (mathematics and writing). They emphasize the specific performance methods used in the assessments rather than the range of performances that characterize the underlying domain.

This narrowness of focus sometimes leads to perverse consequences. It appears that KIRIS has led to large swings in exposure to specific subjects from year to year. For example, fourth-grade teachers reduce the amount of class time devoted to mathematics in favor of the subjects tested at their grade level. This cannot be in the best interest of students' long-term mathematical development. In fact, this behavior is short-sighted even from the narrow perspective of the accountability system, because these same students will be tested on mathematics in the fifth grade. Reduced exposure to mathematics in fourth grade is likely to have a negative impact on subsequent achievement. This is particularly true in the case of higher-order mathematical thinking and reasoning skills (e.g., problem solving, mathematical communication) that are cumulative in nature. One would hope that fifth-grade students have learning experiences designed to help them develop these skills during the fourth grade and in prior years, but this may not be the case if fourth-grade teachers are reducing their coverage of mathematics. Finally, the case studies of exemplary teachers that were conducted in conjunction with these surveys identified other negative consequences of the Kentucky accountability system. For example, exemplary teachers organize their

instruction around the timing of assessments (Borko & Elliott, 1999) even though they do not believe that is the best way to teach the subject (Wolf & McIver, 1999).

One possible reaction to these results is to blame teachers for short-sightedness and for not looking beyond the immediate rewards of the system. While there may be some justification for this reaction, teachers should not shoulder the bulk of responsibility for these results. The teachers' actions are consistent with the incentives of the accountability system. They are attending to the aspects of the accountability system that are most salient to their grade level.

Instead, the majority of the blame for these perverse consequences rests with the accountability system itself. The testing and rewards mechanisms are flawed, and they may be leading teachers to adopt undesirable behaviors. This study does not provide enough information to say which aspects of the testing and rewards systems are responsible for which specific teacher behaviors, but we an speculate about the signals sent by various features of the system.

First, the fact that different subjects are tested in different grade levels may encourage teachers to shift the balance of the curriculum in unusual ways. A fifth-grade teacher who believes that science instruction is as important as mathematics instruction, may nevertheless spend less time on science because only mathematics is an accountability subject in fifth grade.

Second, because the system uses a limited number of measures, teachers may begin to narrow their focus to the tests rather than the domains they are designed to assess. People often fail to remember that tests are just indirect indicators of likely performance in a broad range of demand situations. This may be less of a problem in Kentucky then in many other states because KIRIS uses a range of open-response measures. However, there are still substantial differences between demands of the test items and the performance standards.

Third, because the system focuses on milepost grade levels, it may deflect attention from the cumulative nature of education and imposes undue pressure on teachers in the accountability grades. A student's performance in fourth grade is a function of his or her educational experience since Kindergarten (as well as the myriad of influences outside of the educational system). Yet, the accountability system ignores performance in first, second and third grade. By

focusing on selected grade levels, the system does not signal effectively to other grades. It also distributes the psychological burden unevenly among teachers.

Fourth, because the system establishes very high standards for student performance, teachers may be encouraged to "game" the system. A teacher whose students begin the year unprepared to perform on the accountability assessment, may feel that his or her only option is to work hard on the narrowest definition of the assessment domain—that which has been covered on the assessment in previous years—and hope that some ground can be retaken.

Fifth, because the stakes associated with school scores are so high, teachers may pay greater attention to the test results than to the underlying standards. Teachers who feel undue pressure to raise test scores may let the test specifications and scoring guides unduly influence the curriculum. Stecher and Mitchell (1995) found such behavior among Vermont teachers who were trying to improve student scores on the Vermont mathematics portfolio assessment, and they labeled it "rubric-driven instruction."

If we could identify specific features of the accountability system that led to undesirable consequences, it might be possible to re-design the system to reduce or eliminate them. Although this study was not designed to identify such factors, we can speculate about the aspects of the accountability system that might have promoted the undesirable behaviors we identified.

The problem of shifts in curriculum emphasis from one year to the next might be alleviated by reducing or eliminating milepost testing, i.e., by testing in more grade levels and by testing more subjects per grade level. The more general problem of curriculum narrowing could be addressed in a number of ways. One approach would be to broaden the content of the assessment and to use multiple assessment formats, including constructed response measures. Matrix sampling, a design already in use in Kentucky, could also help to broaden the range of tasks asked of students without increasing the testing burden. (Matrix sampling involves developing larger item pools so teachers do not focus on individual questions but on the skills they are designed to measure.) Changing the content and format of the assessments from year to year may be another promising approach because it sends signals to teachers that the best way to prepare students for the assessment is to focus on the broadest definition of the domain.

However, all these potential solutions involve tradeoffs. Testing in more grades and more subjects adds considerable costs to the system and considerable burdens to teachers and students. The burden is increased if the testing program includes performance assessments in addition to multiple choice tests. Matrix sampled tests typically are not designed to provide valid scores on individuals, which is an important element of accountability in some jurisdictions. Changing content and format from year to year reduces the ability to make comparisons across years, as well as increasing the cost of the assessment. In addition, annual changes create a level of uncertainty that may make teachers uncomfortable and reduce their support for the system.

Another option that might lessen the negative effects of test-based accountability is to reduce the stakes associated with scores. Unfortunately, we have found in other research that there may be no "low stakes" accountability tests. Teachers in a state with public reporting but no financial incentives responded just as strongly to the tests as teachers in a state with public reporting and high stakes (e.g., financial rewards and sanctions) (Koretz et al., 1996).

Reducing the standard for successful performance might also place less pressure on teachers and lead to less distortion of curriculum. High standards were implemented in part as a reaction to the minimum competency tests of the 1980s, a reform that failed to achieve the desired improvements in education. However, it may be the case that we replaced standards that were too low with standards that are too high to serve as good targets for instruction. Even in Kentucky where improvement targets are incremental (i.e., they are based on a school's past KIRIS scores), the demand for continual improvement may quickly lead to targets that are difficult to achieve by educationally sound means. Once teachers begin to feel that they goals are beyond reasonable reach, they may opt to focus narrowly on improving test scores.

Any test-based accountability system will have some unintended and perhaps undesirable consequences. This raises the question of whether test-based accountability on its own is an effective strategy for educational reform? At this point in time, the answer is uncertain. Clearly, accountability systems like the Kentucky system are powerful tools to focus education on particular skills and content. However, the generalizability of score increases has yet to be clearly demonstrated.

One positive sign comes from the 1998 NAEP reading assessment which showed increases is student achievement in Kentucky that were not paralleled in the nation as a whole. If NAEP assessments in other subjects show similar trends then it may be that the accountability system is beginning to pay dividends in Kentucky and the negative effects of the accountability system, such as those demonstrated here, may be concluded to be necessary downsides in an effective program. If, on the hand, future results from assessments such as NAEP fail to indicate that performance is improving in Kentucky in subjects other than reading, the conclusion may be that narrowing of curriculum to focus on what is tested to the exclusion of other important content is producing results that do not generalize beyond the accountability assessment.

Overall, this study shows that teachers have made changes in their classroom practices in response to the state standards and to the KIRIS assessments. While many of the changes are positive and will help students achieve the broader goals of the educational system, there are also perverse effects, such as unjustified shifts in curriculum. States need to attend to such consequences if they opt to implement test-based accountability. They also need to look for ways to design accountability systems that minimize unwanted effects of the type reported here. One important step states can take is to study the consequences of their testing systems as rigorously as they study the reliability and validity of their test scores. Such research would improve our understanding of the features of accountability systems associated with desirable and undesirable practices.

References

Association of California School Administrators. (1996). *Statewide academic standards: Doing it right*. Sacramento: Author.

Bond, L. A., Braskamp, D., & van der Ploeg, A. (1995). *State student assessment programs database: School year 1994-95*. Oakbrook, IL: North Central Regional Educational Laboratory.

Borko, H., & Elliott, R. (1999). Hands-on pedagogy versus hands-off accountability: tensions between competing commitments for exemplary mathematics teaches in Kentucky. *Phi Delta Kappan, 80*(5), 394-400.

Hambleton, R. K., Jaeger, R. M., Koretz, D., Linn, R. L., Millman, J., & Phillips, S. E. (1995). *Review of the measurement quality of the Kentucky Instructional Results Information System, 1991-1994*. Frankfort: Kentucky General Assembly.

Hill, P. T., & Bonan, J. J. (1991). *Decentralization and accountability in public education*. R-4066-MCF/IET. Santa Monica: RAND.

Kellaghan, T., & Madaus, G. (1991). National testing: Lessons for America from Europe. *Educational Leadership, 49*(3), 87-93.

Kentucky Department of Education. (1994). *Kentucky's learning goals and academic expectations*. Frankfort: Author.

Kentucky Department of Education. (1995). *Transformations: Kentucky's curriculum framework*. Frankfort: Author.

Kentucky Department of Education. (1996). *Core content for assessment*. Frankfort: Author.

Klein, S. P., McCaffrey, D., Stecher, B., & Koretz, D. (1995). The reliability of mathematics portfolio scores: Lessons from the Vermont experience. *Applied Measurement in Education, 8*(3), 243-260.

Koretz, D. M. (1998). Large-scale portfolio assessments in the U.S.: Evidence pertaining to the quality of measurement. *Assessment in Education, 5*(3), 309-334.

Koretz, D. M., & Barron, S. I. (1998). *The validity of gains in scores on the Kentucky Instructional Results Information System (KIRIS)*, MR-1014-EDU. Santa Monica: RAND.

Koretz, D. M., Barron, S. I., Mitchell, K. J., & Stecher, B. M. (1996). *Perceived effects of the Kentucky Instructional Results Information System (KIRIS).* MR-792.PCT/FF. Santa Monica: RAND

Koretz, D. M., Klein, S., McCaffrey, D., & Stecher, B. (1993). Interim report: *The reliability of Vermont portfolio scores in the 1992-93 school year* (CSE Tech. Report 370). Los Angeles: UCLA National Center for Research on Evaluation, Standards and Student Testing.

Koretz, D. M., Linn, R. L., Dunbar, S. B., & Shepard, L. A. (1991). *The effects of high-stakes testing on achievement: Preliminary findings about generalization across tests.* In R. L. Linn (Chair), *Effects of high-stakes testing on instruction and achievement.* Symposium presented at the annual meeting of the American Educational Research Association and the National Council on Measurement in Education, Chicago.

Koretz, D. M., Mitchell, K. J., Barron, S. I., & Keith, S. (1996). Final report: *Perceived effects of the Maryland School Performance Assessment Program* (CSE Tech. Report 409). Los Angeles: UCLA National Center for Research on Evaluation, Standards and Student Testing.

Koretz, D. M., Stecher, B., Klein, S., & McCaffrey, D. (1994). The Vermont portfolio assessment program: Findings and implications. *Educational Measurement: Issues and Practice, 13*(3), 5-16.

Linn, R. L. (1998). *Standards-based accountability: Ten suggestions.* Los Angeles: UCLA CRESST.

Linn, R. L., Graue, M. E., & Sanders, N. M. (1990). Comparing state and district results to national norms: The validity of the claims that "everyone is above average." *Educational Measurement: Issues and Practice, 9*(3), 5-14.

Office of Technology Assessment. (1992). *Testing in American schools, asking the right questions.* Washington, DC: Congress of the United States.

Resnick, L. B., & Resnick, D. P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B. G. Gifford & M. C. O'Connor (Eds.), *Changing assessment: Alternative views of aptitude, achievement, and instruction* (pp. 37-75). Boston: Kluwer Academic Publishers.

Shepard, L. A., & Dougherty, K. C. (1991). *Effects of high-stakes testing on instruction.* In R. L. Linn (Chair), The effects of high stakes testing, symposium presented at the annual meeting of the American Educational Research Association and the National Council on Measurement in Education, Chicago, IL.

Smith, M. L., & Rothenberg, C. (1991). Unintended consequences of external testing in elementary schools. *Educational Measurement: Issues and Practice, 10*(4), 7-11.

Stecher, B. (1998). The local benefits and burdens of large-scale portfolio assessment. *Assessment in Education, 5*(3), 335-351.

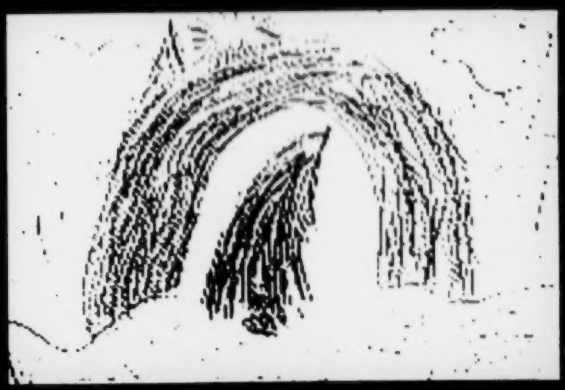
Stecher, B. M., & Herman, J. L. (1997). Using portfolios for large-scale assessment. In G. D. Phye (Ed.), *Handbook of classroom assessment*. Orlando: Academic Press.

Stecher, B. M., & Mitchell, K. J. (1995). *Portfolio driven reform; Vermont teachers' understanding of mathematical problem solving*, CSE Technical Report 400, Los Angeles, CA: UCLA/National Center for Research on Evaluation, Standards and Student Testing (CRESST).

Stecher, B. M., Barron, S. I., Kaganoff, T., & Goodwin, J. (1998). *The effects of standards-based assessment on classroom practices: Results of the 1996-97 RAND survey of Kentucky teachers of mathematics and writing* (CSE Technical Report 482). Los Angeles: UCLA/National Center for Research on Evaluation, Standards and Student Testing (CRESST).

Washington State Commission on Student Learning. (1997). *Essential academic learning requirements: Technical manual*. Olympia: Author.

Wolf, S. A., & McIver, M. C. (1999). When progress becomes policy: the paradox of Kentucky state reform for exemplary teachers. *Phi Delta Kappan, 80*(5), 401-406.



UCLA Graduate School of Education & Information Studies